Stability selection:

enhanced variable selection and network models

Session 2/3

Barbara Bodinier

Computational Epidemiology
MSc Health Data Analytics and Machine Learning

4 March 2021

Imperial College
London

Refresher on performance metrics

- Contingency table:

	Selected	Not selected
Contributing	True Positive (TP)	False Negative (FN)
Not contributing	False Positive (FP)	True Negative (TN)

- Sensitivity = recall: $TPR = r = \frac{TP}{TP+FN}$
- Specificity: $TNR = \frac{TN}{TN+FP}$
- Accuracy: $a = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $p = \frac{TP}{TP+FP}$

$\Rightarrow$ Which one would you use?

Refresher on performance metrics

- Contingency table:

	Selected	Not selected
Contributing	True Positive (TP)	False Negative (FN)
Not contributing	False Positive (FP)	True Negative (TN)

- Sensitivity = recall: $TPR = r = \frac{TP}{TP+FN}$
- Specificity: $TNR = \frac{TN}{TN+FP}$
- Accuracy: $a = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $p = \frac{TP}{TP+FP}$

$\Rightarrow$ In network models, we expect sparse models, i.e. a large number of negatives (not selected) compared to the positives (selected)

$\Rightarrow$ Specificity/accuracy would be very high with very few (or no) selected features

$\Rightarrow$ Same as in a classification problem with much more controls than cases

Refresher on performance metrics

- Contingency table:

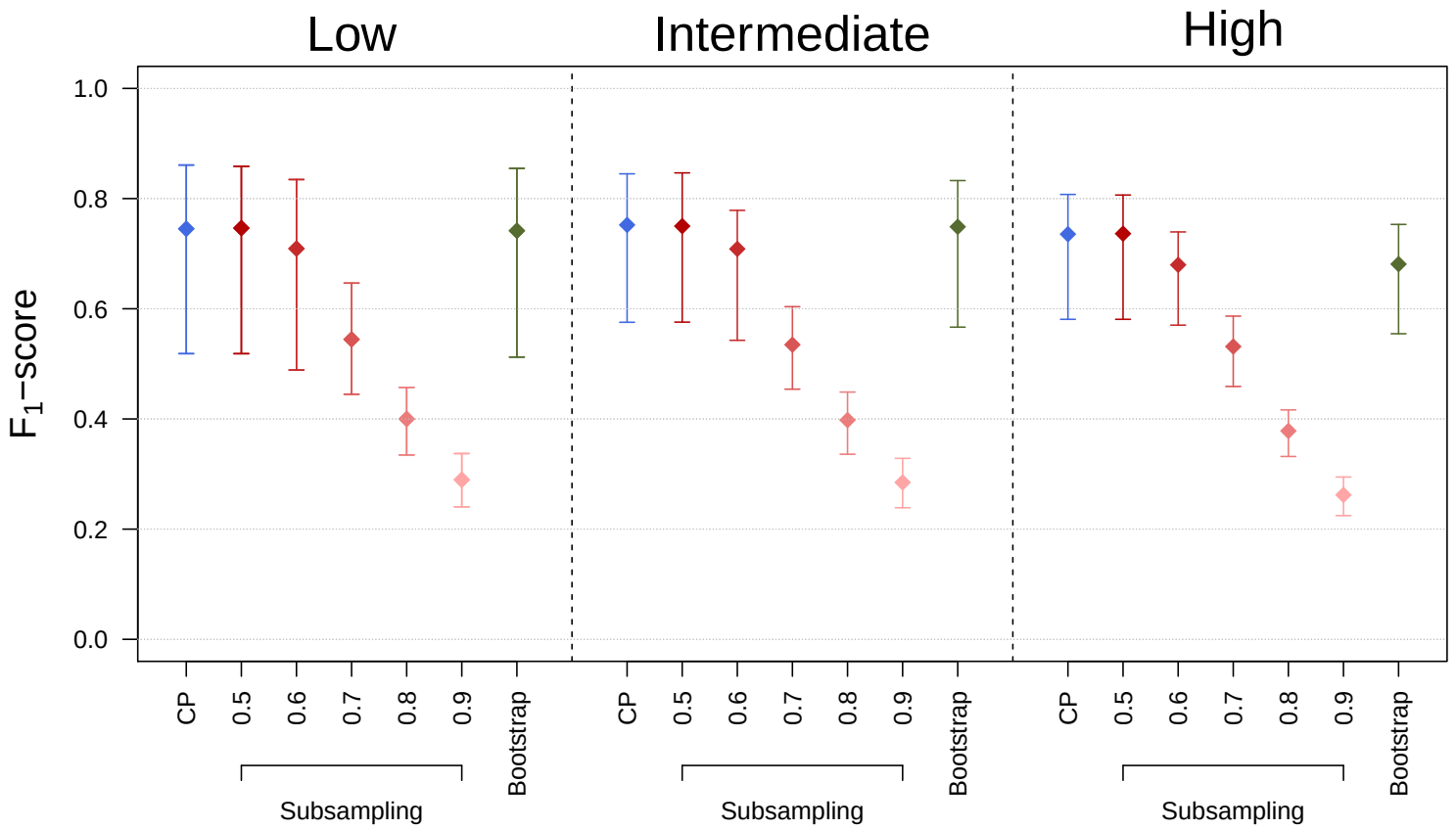
	Selected	Not selected
Contributing	True Positive (TP)	False Negative (FN)
Not contributing	False Positive (FP)	True Negative (TN)

- Sensitivity = recall: $TPR = r = \frac{TP}{TP+FN}$
- Specificity: $TNR = \frac{TN}{TN+FP}$
- Accuracy: $a = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $p = \frac{TP}{TP+FP}$

$\Rightarrow$ Evaluation of the models will be done in terms of precision/recall
$\Rightarrow$ Overall performance can be summarised in one metric: $F_1 = \frac{2 \times p \times r}{p+r}$

Simulation study: the choice of resampling approach

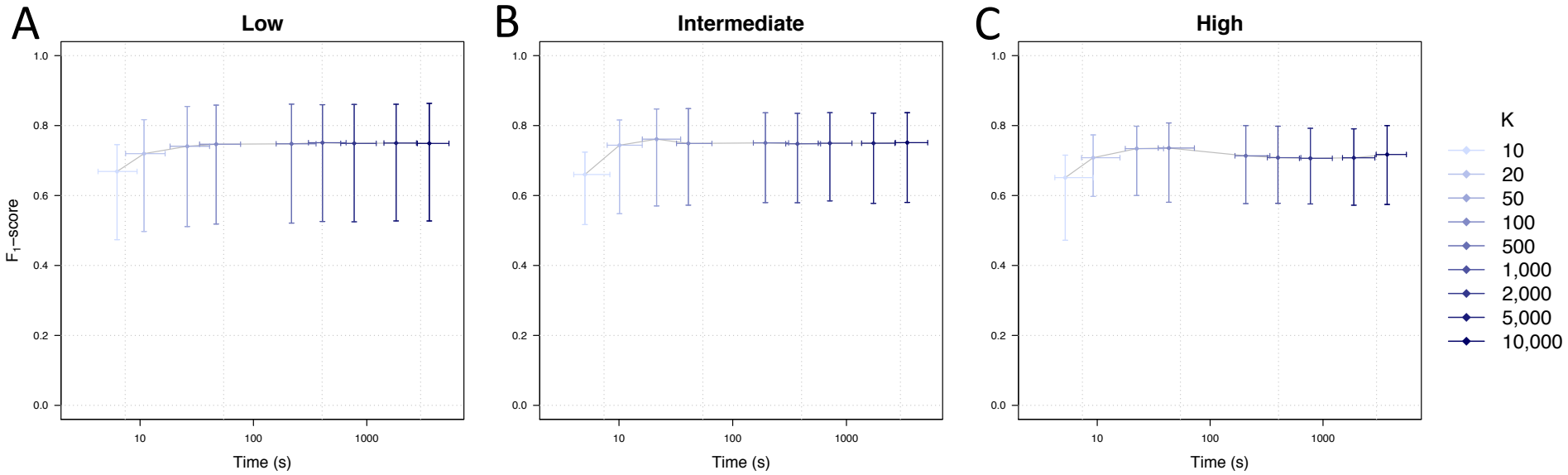
- Comparison of stability selection results with different resampling approaches



$\Rightarrow$ Boostrap/subsampling of 50% of the observations gives best results in terms of F_1-score

Simulation study: the number of iterations

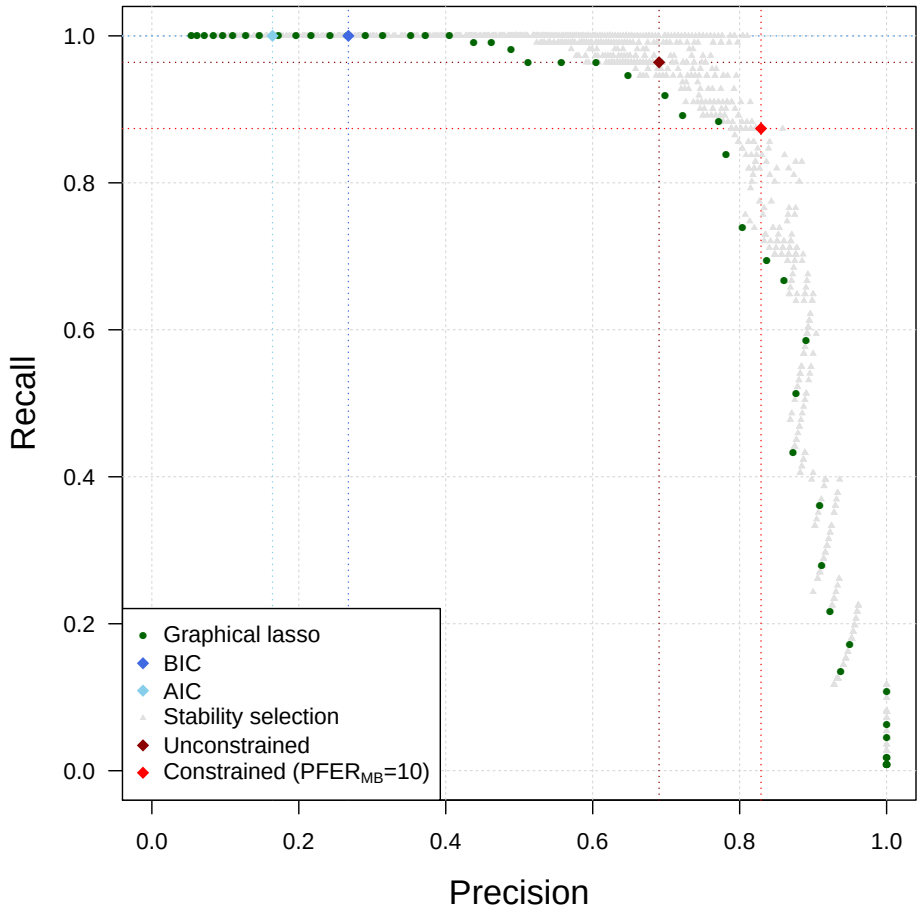
- Comparison of stability selection results with different numbers of iterations



$\Rightarrow$ Similar performances with $\geq$ 50 iterations for a network estimation problem with p = 100 nodes and n = 200 (low dimension), n = 100 (intermediate) and n = 50 (high) observations

Simulation study: comparison with non-stability models

- Precision-recall plot and comparison with non-stability approaches

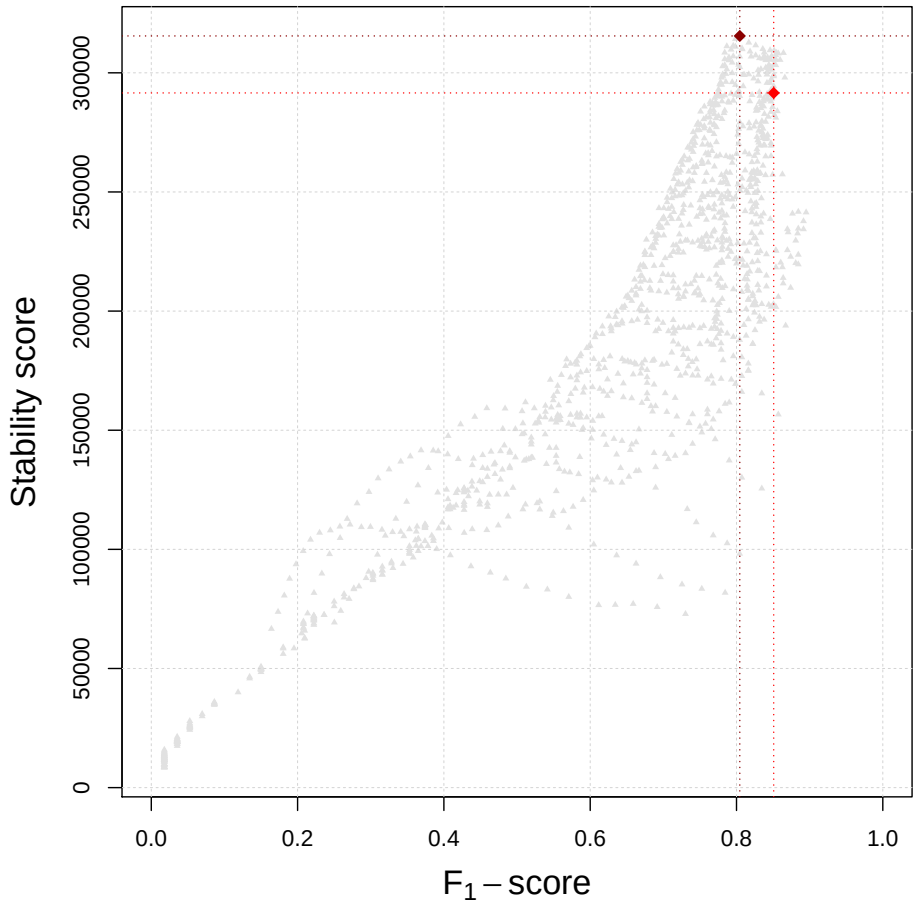


$\Rightarrow$ Regardless of the parameters, stability selection models (grey) are out-performing graphical lasso models (dark green)

$\Rightarrow$ Worth moving to stability-enhanced models

Performance of the stability score

- Stability score as a function of model performance (as measured by the F_1-score)

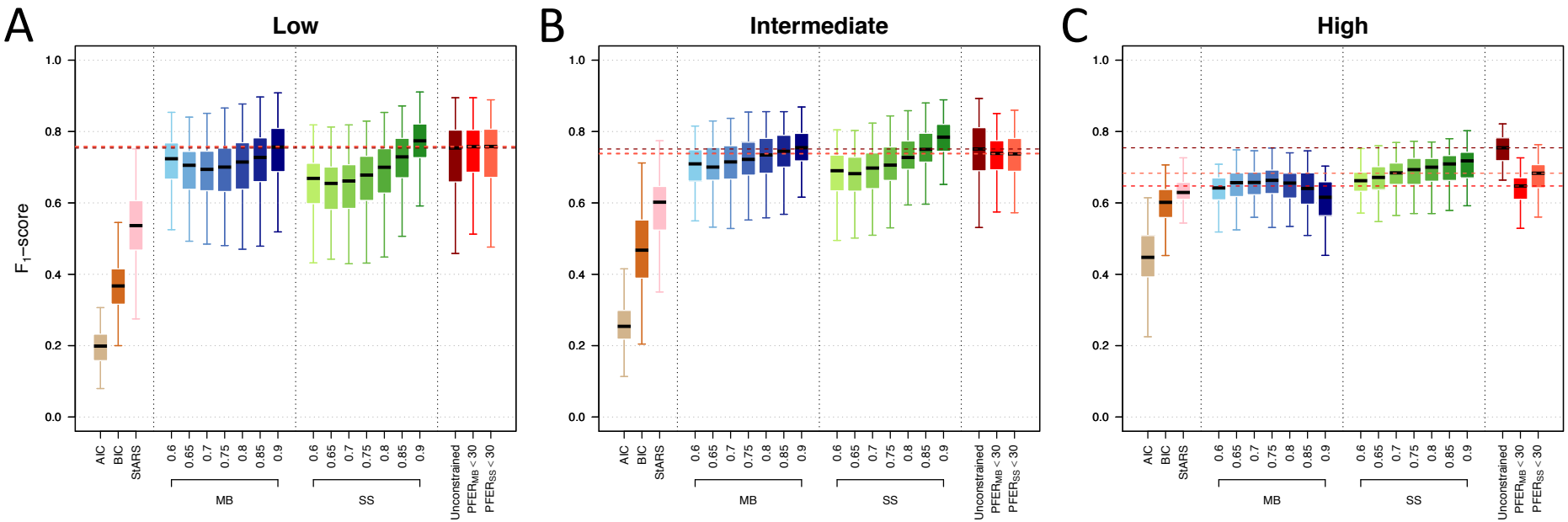


$\Rightarrow$ Stability score is increasing with model performance

$\Rightarrow$ Stability score is a relevant criterion for calibration

Comparison with existing approaches

- F_1-score of our model (red) is compared to that of existing approaches in different dimensionality settings



$\Rightarrow$ Our model out-performs the graphical lasso calibrated by AIC or BIC

$\Rightarrow$ Comparable and generally better performance than the original MB or SS procedures with one arbitrarily set parameter

Real data application: metabolomic markers of lung cancer risk

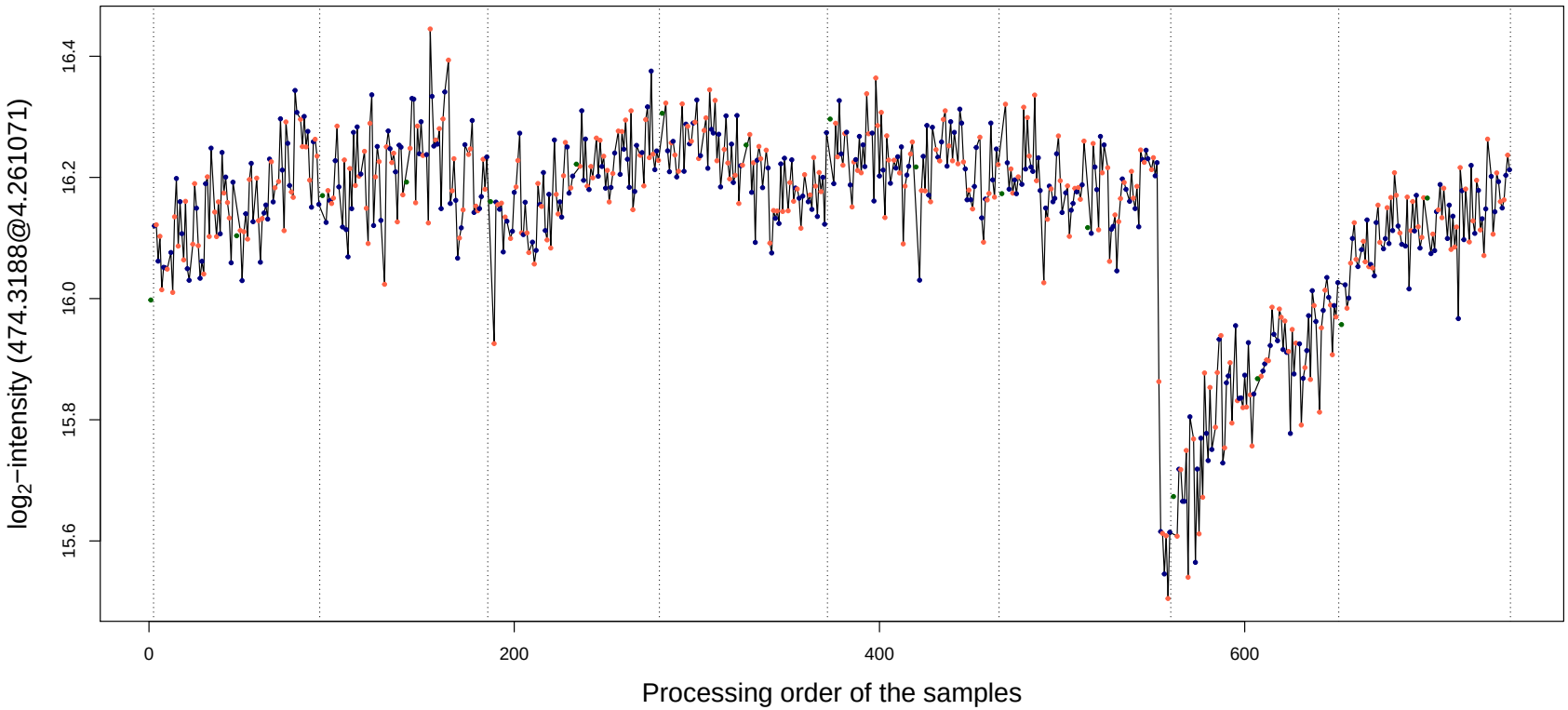
- Prospective lung cancer study on 648 participants from EPIC Italy (N=382) and NOWAC (N=266)
- Untargeted metabolomics measurements on two platforms:
  - N=5,011 features in positive ionisation mode
  - N=11,767 features in negative ionisation mode
- Technical issue: drop in intensity for plates 6, 7 and 8 for positive mode

- Data preparation:
  1. $\log_2$-transformation of the data
  2. Exclusion of the features with more than 50% of missing in samples
  3. Correction for technical confounding by extracting the residuals and intercept from linear mixed models with random intercepts for center and plate and random slope of position by plate applied on the 643 samples (complete cases)
  4. Background subtraction

$$\Rightarrow \text{N=1,516 (positive) and N=2,452 (negative)}$$

Accounting for technical confounding

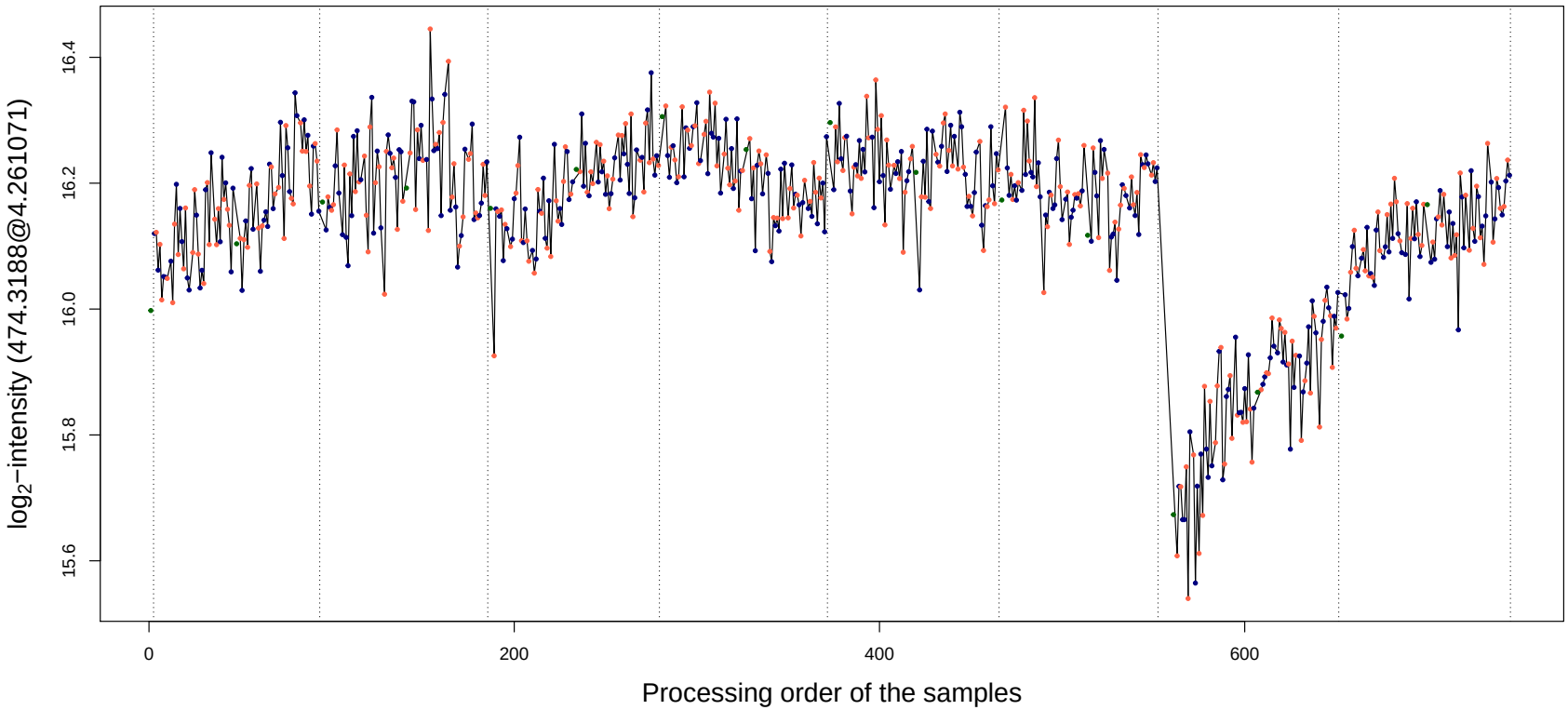
- Intensity of reference feature (474.3188@4.261) in order of processing
- Feature is expected to have the same levels in all samples
- Cases are shown in red, controls in blue and blanks in green



$\Rightarrow$ Clear visualisation of the drop in intensity just before the end of plate 6

Intensity of reference feature (positive mode)

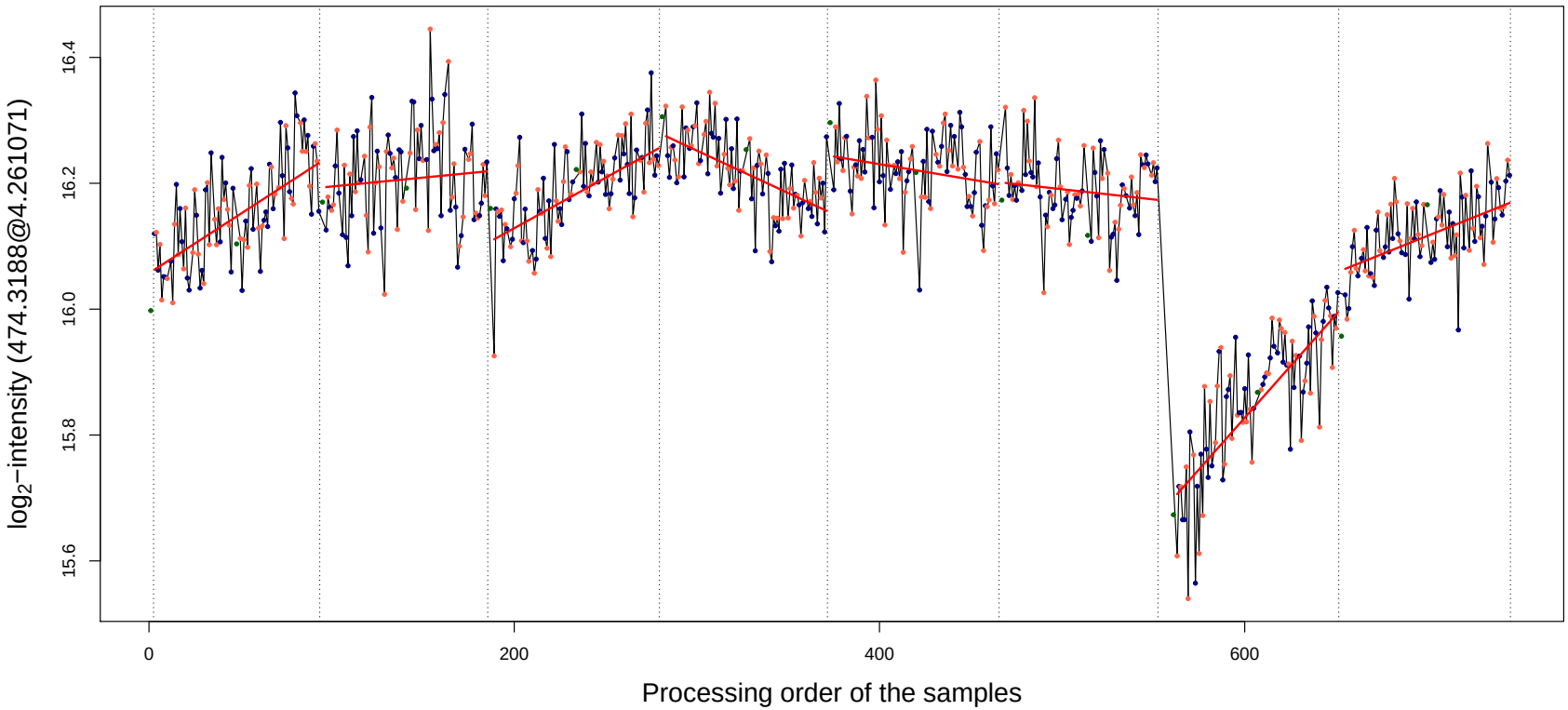
- After removing the N=7 samples after the drop on plate 6:



$\Rightarrow$ Variability in intensity within plates, random intercept on plate is not enough

Intensity of reference feature (positive mode)

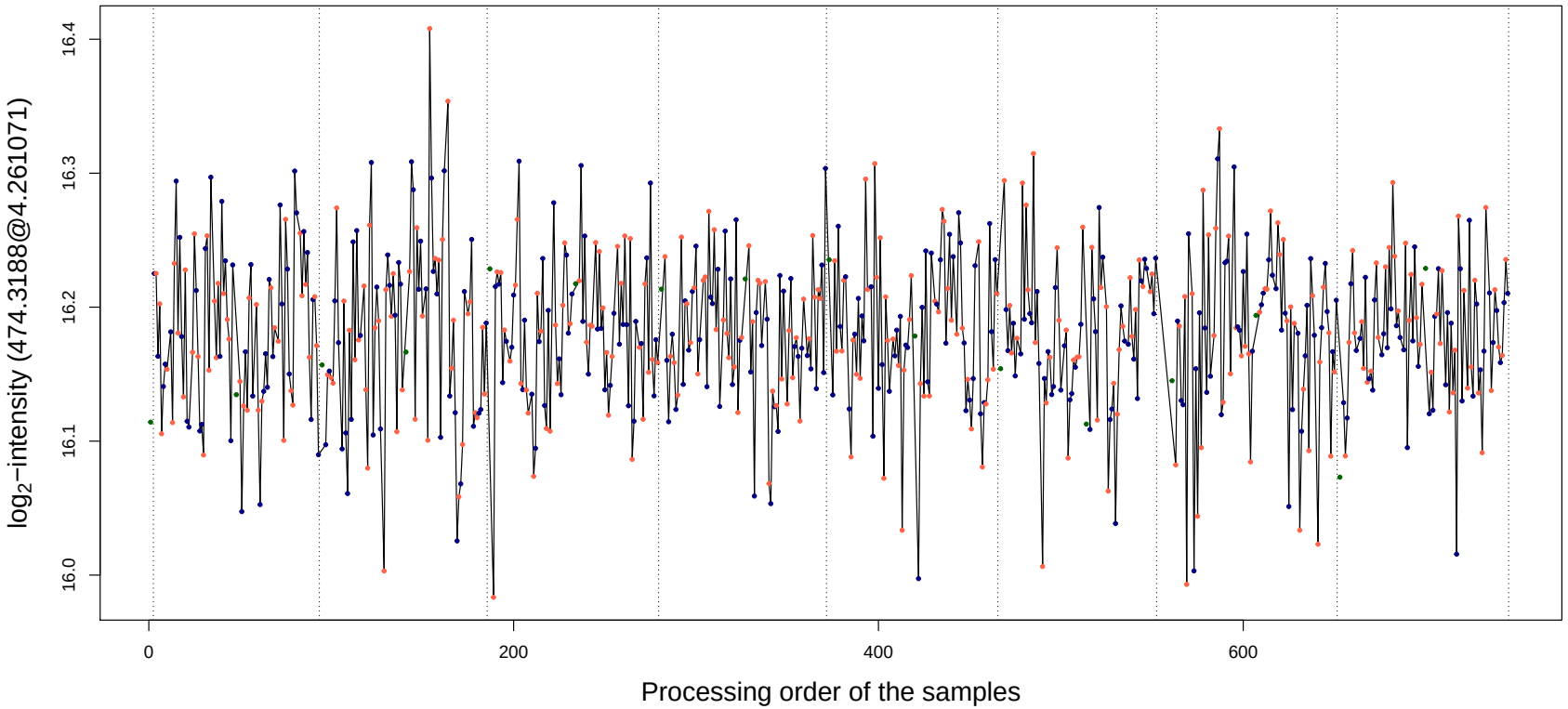
- Linear mixed model with random intercept on plate and random slope on the position (order of processing)



⇒ Linear mixed models detect the plate effect and trend in intensity within plates

Intensity of reference feature (positive mode)

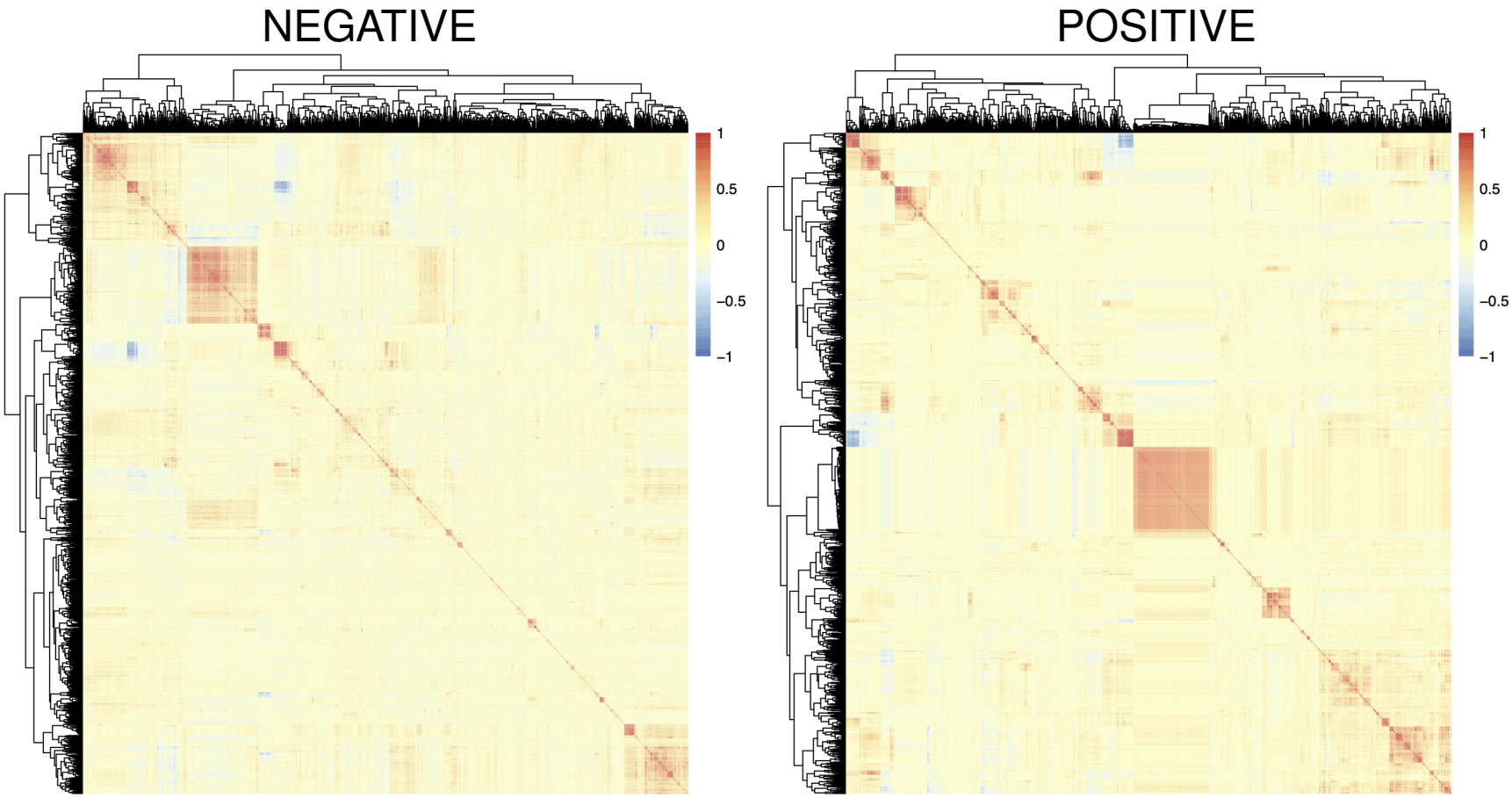
- Extracting the intercept (to keep the scale) and residuals from the linear mixed model:



$\Rightarrow$ Levels are corrected for the effect of plate and linear changes in intensity

Correlation between metabolic features

- Pearson's correlation between log-transformed levels:



⇒ Patterns of strong within-mode correlations (potentially between features from the same molecule)

⇒ Multivariate models need to accommodate these strong correlations

Clustering-summary of the data

- Two-step strategy:
  1. Clustering to identify groups of metabolomics features that are strongly correlated

     $\Rightarrow$ Agglomerative hierarchical clustering maximising the Euclidian distance between PC_1-scores summarising the clusters
  2. Summary of each of the identified clusters as one variable

     $\Rightarrow$ Each cluster is summarised by the scores from the first PC of a Principal Component Analysis applied on the features grouped together

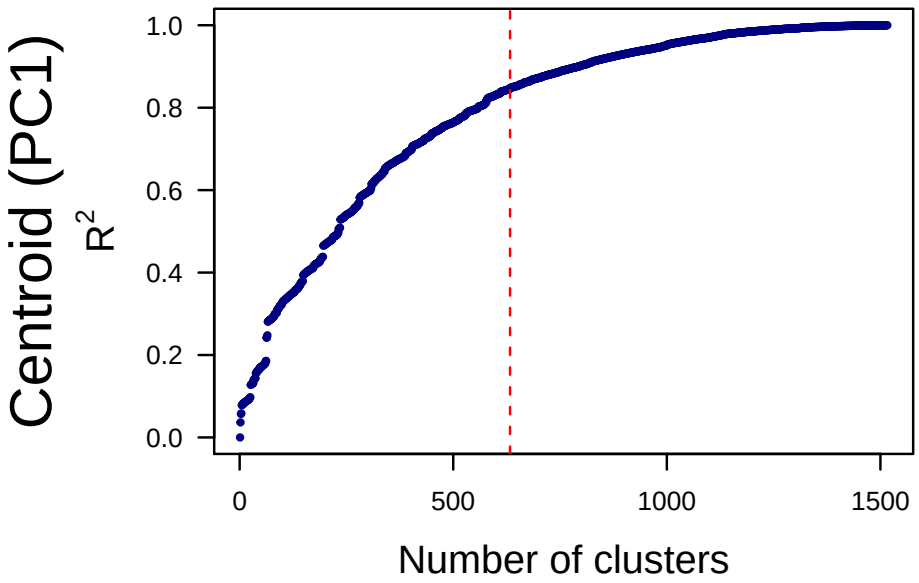
$\Rightarrow$ Destroy the strong correlations causing problems in multivariate models

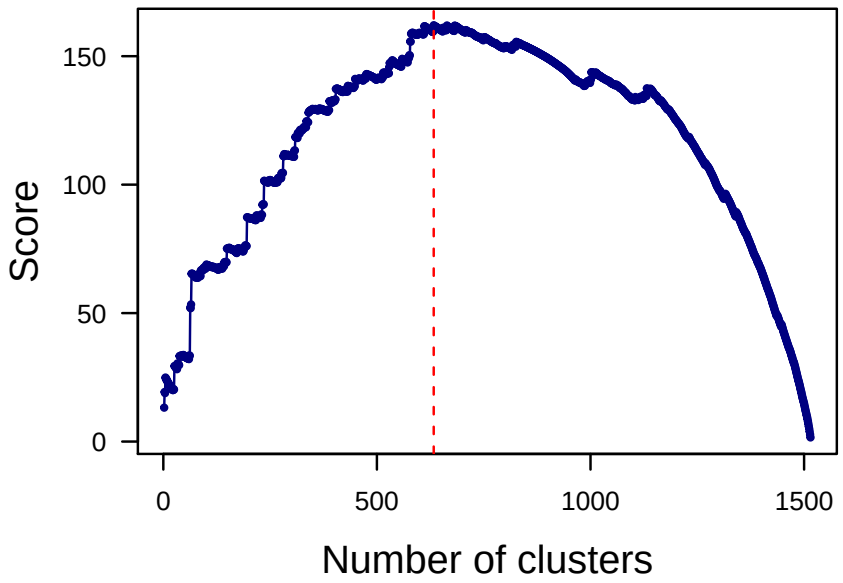
$\Rightarrow$ Simplify the data by excluding redundant features

$\Rightarrow$ Need to calibrate the number of clusters such that there is a good trade-off between the compactness (within-cluster correlation) and separation (between-cluster distance)

Clustering-summary of the positive mode data

- Calibration of the number of clusters to optimise trade-off in compactness and separation
- Visualisation of the proportion of explained variance with different numbers of summarised clusters

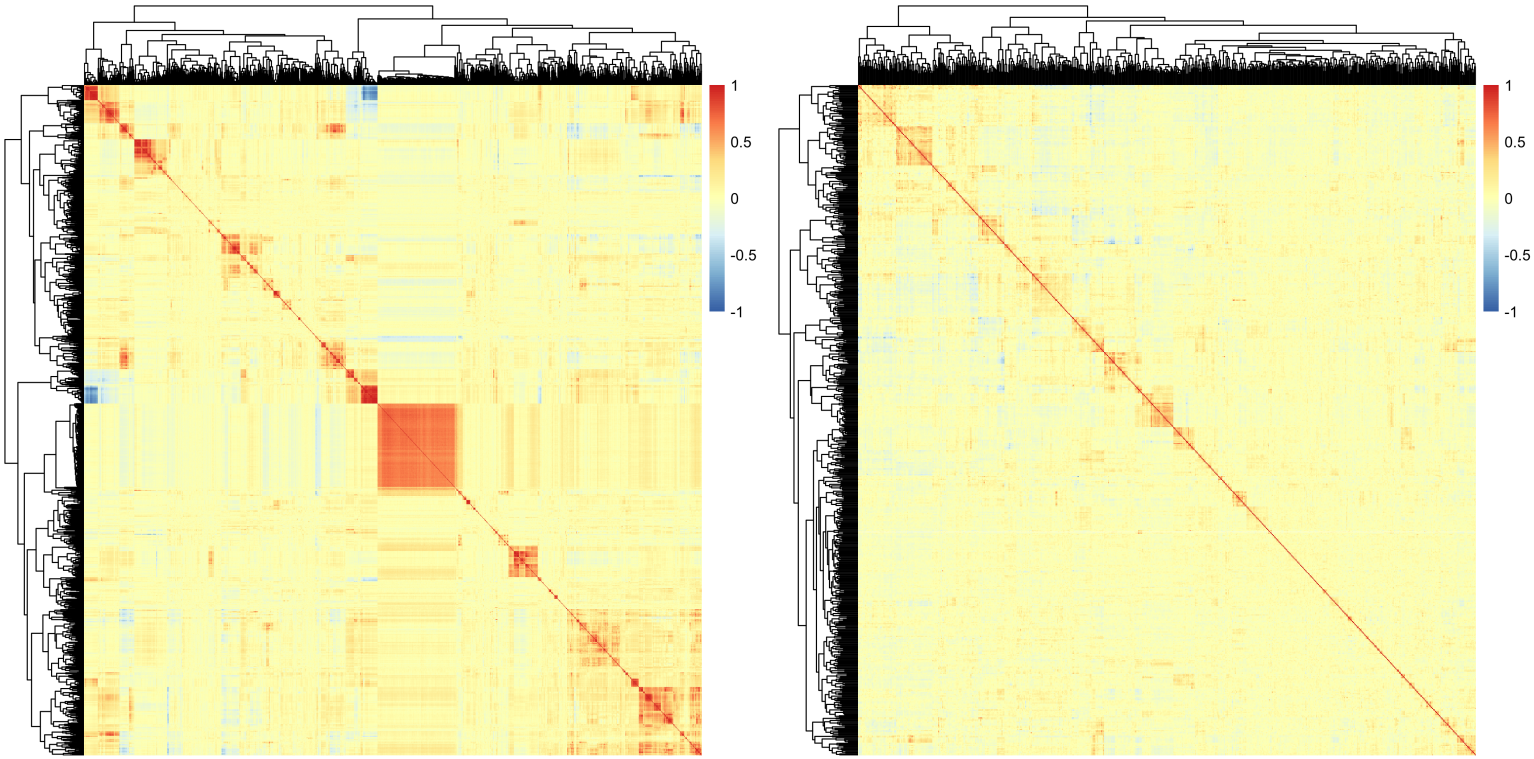




$\Rightarrow$ Summarised data with 633/1516 features (one variable per cluster) explains 85% of the total variability in original data

Visualisation of summarised data (positive mode)

- Heatmap of correlations in the original (left) and summarised data using PC1-scores (right)

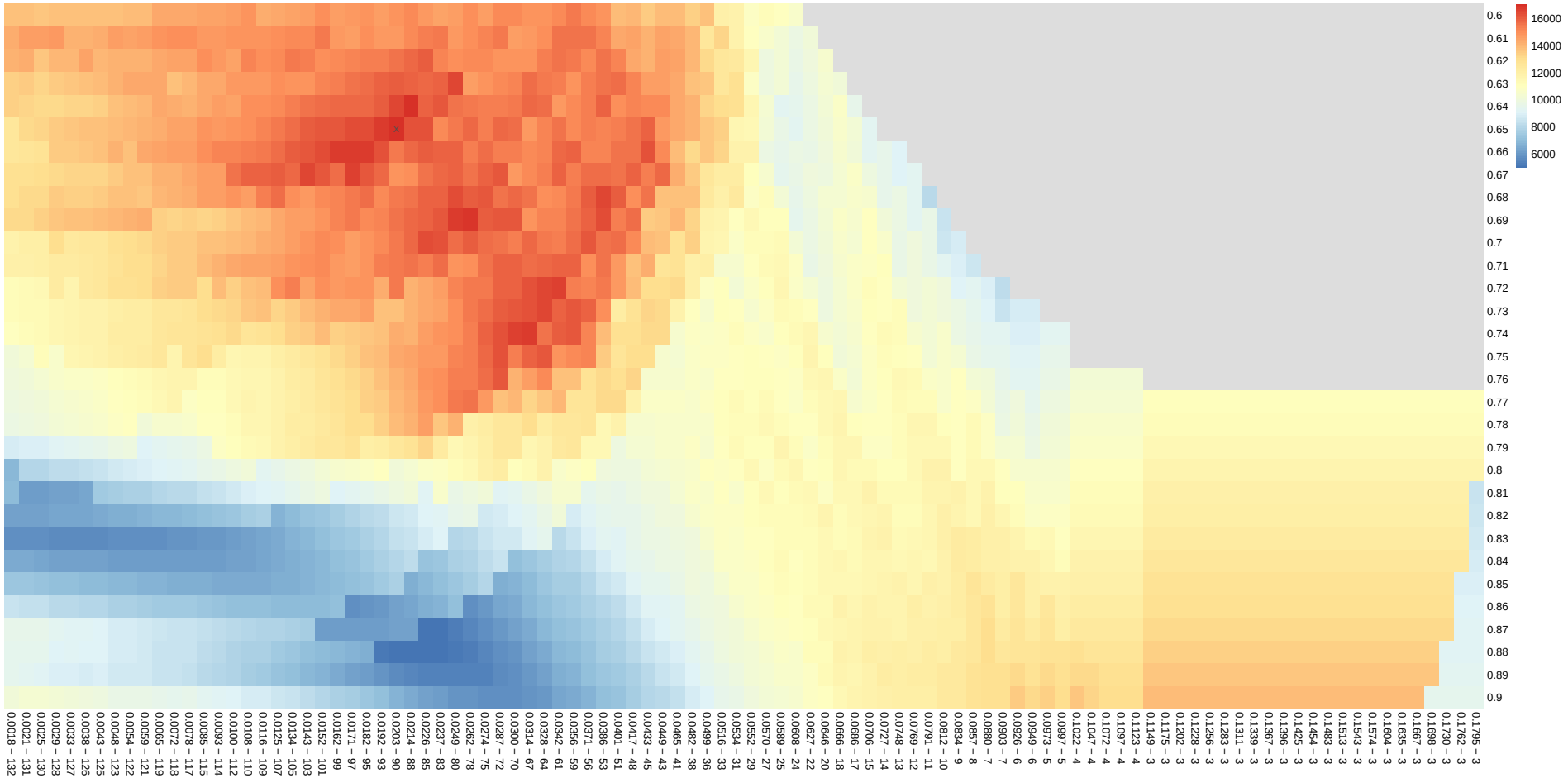


$\Rightarrow$ Strongest correlations have been removed

$\Rightarrow$ Strongest correlation of 0.77 in the summarised data

Stability selection logistic-LASSO

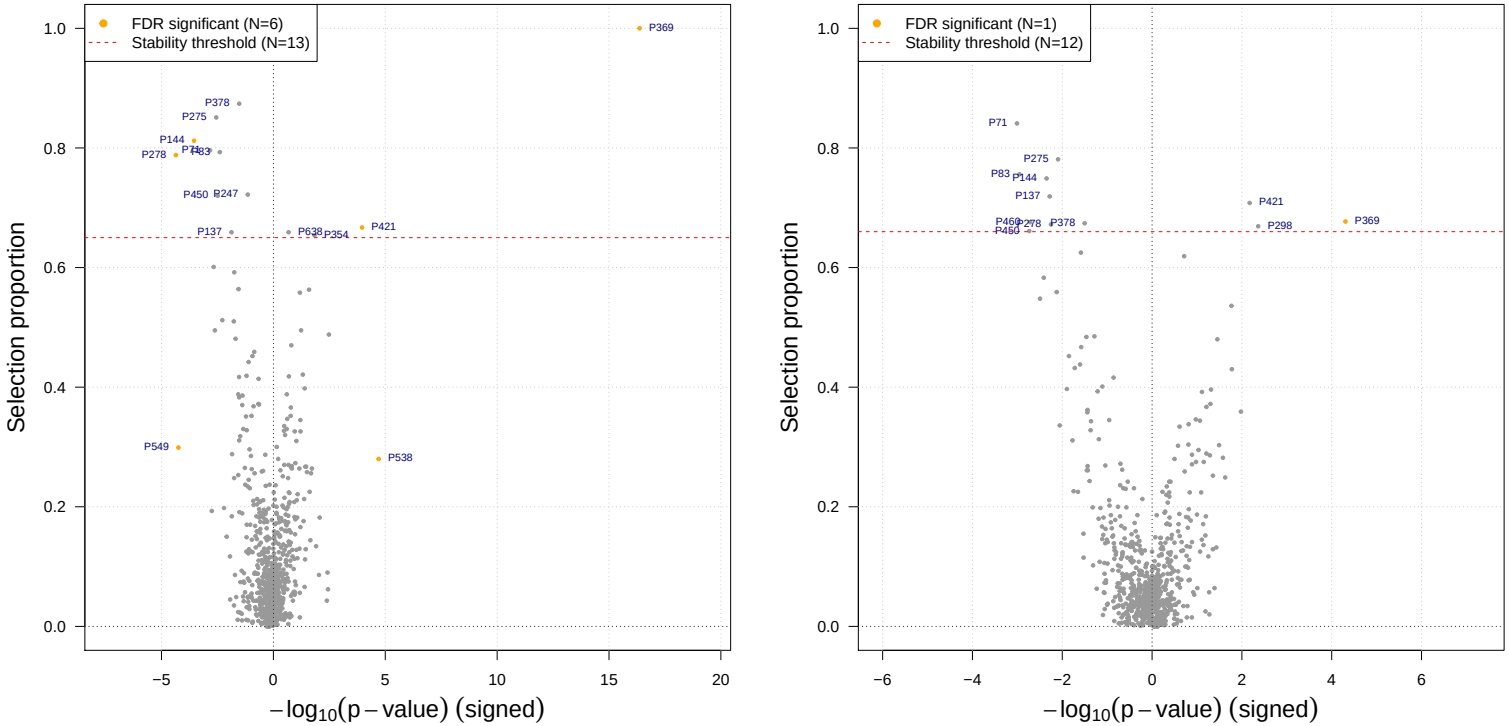
- Calibration of the stability selection model applied on 633 features
- Models are adjusted on age and BMI



$\Rightarrow$ Calibrated $\lambda = 0.0203$ and $\pi = 0.65$

Stability selection logistic-LASSO

- Comparison with univariate models: base model and model further adjusted on packyears



$\Rightarrow$ Four of the six univariate hits are selected in the base stability model

$\Rightarrow$ Additional signals in model adjusted on smoking (N=11)

$\Rightarrow$ Strong attenuation of the selection proportion and association for P369 upon adjustment on smoking